DOCUMENT RESUME

ED 307 339 TM 013 527

AUTHOR Linacre, John M.
TITLE Objectivity for Judge-Intermediated Certification Examinations.
PUB DATE Mar 89
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Evaluators; *Interrater Reliability; *Latent Trait Theory; *Licensing Examinations (Professions); Models; Testing Problems
IDENTIFIERS *Fairness; *Objectivity; Rasch Model; Stochastic Approximation Method

ABSTRACT

An accepted criterion for gauging the fairness of examinees' scores, derived from judge-awarded ratings, has been the size of the correlation between the judges and the inter-rater reliability. Various means of achieving inter-rater reliability were reviewed, and a model to measure inter-rater reliability is forwarded. Both theoretical and practical considerations mandate that perfect inter-rater reliability can never be achieved. A stochastic element always remains. Objective measurement of examinees, freed from the severity of the judges and the definition of the rating scale, can be obtained by capitalizing on the stochastic nature of ratings. The resulting measurement model is of the type developed by Rasch. Examples of the model are provided. (TJH)

Objectivity for

Judge-intermediated Certification Examinations

John M. Linacre

M E S A

University of Chicago

5835 S. Kimbark Avenue

Chicago Il 60637

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, March 31, 1989.

Abstract:

An accepted criterion for gauging the fairness of examinees' scores, derived from judge-awarded ratings, has been the size of the correlation between between the judges, the inter-rater reliability. Both theoretical and practical considerations mandate that perfect inter-rater reliability can never be achieved. A stochastic element always remains. Objective measurement of the examinees, freed from the severity of the judges and the definition of the rating scale, can be obtained ' capitalizing on the stochastic nature of ratings. The resulting measurement model is of the type developed by Rasch, and examples of the model are given.

Key-words:

Judging, Rating scales, Objectivity, Rasch measurement, Reliability.

I. Introduction:

The use of judges to assess the performance of examinees is viewed as undesirable but, on occasion, necessary. The reasons for this reluctance to use judges is clear. As Braun has written of essay examinations, "large numbers of graders must be trained and supervised and the maintenance of uniform standards across graders and over many days often becomes problematic. An immediate consequence is that the reliability of the scores is often substantially less than unity" (Braun 1988 p.1). The problem of judge training is directly related to the problem of inter-rater reliability. How the goal of a perfect test is perceived affects how judges are trained towards that goal.

The quest for perfect reliability in judge ratings of examinees is pervasive in the literature of judging and has motivated the implementation of techniques to improve judging quality, such as clearer definition of the categories of rating scales and more precise instructions to the judges. But, like the search for El Dorado, the quest for reliability is ultimately doomed to failure. This does not mean, however, that the ratings of judged performances cannot be just as trustworthy as the responses to multiple-choice questions. What it does mean is that emphasis must be placed, not on numerical agreement between the ratings of the different judges (reliability), but on agreement between the intentions of the judges. This change of emphasis enables the construction of measures for the examinees which are independent, in meaning, of the particular judges who rated each performance and so are, colloquially speaking, "fair", or technically speaking, "objective".

II. Why is the quest for reliability doomed to failure ?

The ideal judging situation would appear to be that in which all judges agree on every rating of examinees that they share in common. An example of this is shown in Figure 1. Examinees have been given the same ratings by two independent judges. For the purposes of this immediate discussion, the

rating scale is assumed to be an equal interval scale, so that the usual arithmetic operations can be performed on it. This is an assumption in most analysis of rating scales, such as Guilford (1954 p.278-301).

	Examinees 1 2 3 4 5 6
Judge A	4 3 4 2 1 3
Judge B	4 3 4 2 1 3

Figure 1. Perfect agreement in judges ratings of six examinees on some task. The rating scale has 5 categories in ascending order of performance level, 0,4.

Whether perfect agreement is, in fact, the ideal has been questioned by a number of researchers. On the one hand, from the empirical viewpoint, "when two examiners award different marks, the average is more likely to be correct, or nearly correct, than it is when they award the same mark" (Harper 1976 p.262). On the other hand, from the theoretical viewpoint, "it is usually required to have two or more raters who are trained to agree on independent ratings of the same performance. It is suggested that such a requirement may produce a paradox of attenuation associated with item analysis, in which too high a correlation between items, while enhancing reliability, decreases validity" (Andrich 1984).

	Examinees 1 2 3 4 5 6
Judge A	4 3 4 2 1 3
Judge C	3 2 3 1 0 2

Figure 2. Perfect inter-rater reliability between in the ratings of six examinees on some task. Judge C is one score-point more severe than Judge A.

The topic of complete agreement, however, is a moot point, because it cannot be expected to occur in any large-scale examination situation. Given that raters do differ, let us consider the question of perfect judge reliability. Figure 2 gives an example of this, under the same conditions and with the same six examinees as Figure 1.

It can be seen that the ratings given by the judges are perfectly correlated, and so perfectly reliable according to indices based on product-moment correlation. However, according to indices based on nominal agreements in the ratings, such as Cohen's (1960) Kappa, there is no agreement at all. It

can be seen that these judges do agree on the rank order of the examinees, so that to report that they have no agreement at all is clearly misleading. Accordingly we will consider these judges to have perfect inter-rater reliability, but we discern that judge A is one score-point more lenient than Judge C. In this paper, we will consider judges to differ only in leniency/severity because this is usually a large component of the variance of the ratings, and, as Braun suggests as a result of his study, "adjusting scores for the differences between [judges] should improve the reliability" (Braun 1988 p.8).

The possibility of perfect inter-rater reliability of judges whose severity differs by one score point raises the question of what would be the ratings given by a perfectly reliable judge who is only 0.5 score points more severe than Judge A. Figure 3 makes one suggestion. Judge D must accommodate his behavior to the predefined rating scale, and thus his 0.5 point difference is expressed by awarding half the examinees a rating one point lower than Judge A, and the other half the same rating as Judge A. Consequently, two judges, who, in intention, have perfect reliability, are observed to have a correlation coefficient of 0.895.

Let us say that, as a result of some analysis, Judge D has been determined to be 0.5 score points more severe than Judge A, and a correction of 0.5 points is made in all Judge D's ratings. The outcome is shown in Figure 4. The inter-rater reliability has not changed, nor, after rounding to the nearest integer category, has the nominal agreement in categories. The correction for judge severity has made no improvement in the reliability of this set of ratings.

	Examinees					
	1	2	3	4	5	6
Judge A	4	3	4	2	1	3
Judge D	3	3	3	2	0	3

Figure 3. Ratings given by two judges when one judge is 0.5 score-points more lenient than the other.

	Examinees					
	1	2	3	4	5	6
Judge A	4	3	4	2	1	3
Judge D	3.5	3.5	3.5	2.5	0.5	3.5

Figure 4. Ratings given by two judges when a judge severity of 0.5 score-points has been corrected for.

Another judge, E, who is also 0.5 score points more severe that Judge A, now awards his ratings and these are shown in Figure 5. Again, Judge E expressed his severity by awarding half the examinees a rating one score point below that of Judge A, and the other half the same rating as judge A. Again their agreement, in intention, is perfect but their correlation coefficient is 0.895. Now compare Judges D and E, who are both 0.5 score points more severe than Judge A. Their ratings are shown in Figure 6.

We know that both Judge D and Judge E have the same degree of severity and agree, in intention, as to the standard of performance of the examinees, but the constraints of the rating scale have caused them to express it differently. Judge D and Judge E are reported to have a correlation coefficient 0.645, but, from the point of view of an examining board, even this understates the problem. If a rating of 3 or 4 constituted a pass, and 0,1 or 2 constituted a failure, then Judge D passes 4 and fails 2, and Judge E passes 2 and fails 4. Judge E's two passes, however, are reported as maximum scores (4). In traditional analysis, Judges D would be reported as being more severe but generally in agreement with Judge A, but Judge E would be reported as having a marked inversion of "central tendency".

This paradox of lack of reliability has been presented in terms of one judge being 0.5 score points more severe than another. The very same situation arises, however, when one examinee is 0.5 score points less able than another, even when there is a perfect correlation of judge intentions. Indeed, since the process of measurement is based on the concept that there is a continuum of examinee performance, examinees will always be found who perform, for any particular judge, at or near the transition between adjacent categories.

	Examinees 1 2 3 4 5 6
Judge A	4 3 4 2 1 3
Judge E	4 2 4 1 1 2

Figure 5. Further example of ratings given by two judges when one judge is 0.5 score-points more severe than the other.

	Examinees 1 2 3 4 5 6
Judge D	3 3 3 2 0 3
Judge E	4 2 4 1 1 2

Figure 6. Comparison of the ratings awarded by two judges of equal severity.

Even given ideal judges, it is clear that a performance at a level of 2.5 score-points will be awarded 3 points or 2 points with approximately equal frequency. However, with real judges, however well-trained and experienced, what will happen to a performance at a level of 2.49 score-points ? It cannot be expected that the judges will have such precise discrimination that this performance will always be awarded a rating of 2, but never a rating of 3. Indeed we expect a greater frequency of 2's than 3's, but not a very great difference. By extension of the same argument, the situation in Figure 7 can be expected to result. What appeared to be a deterministic decision by judges is revealed to be a probabilistic one. This stochastic element in rating is what dooms the quest for perfect inter-rater reliability.

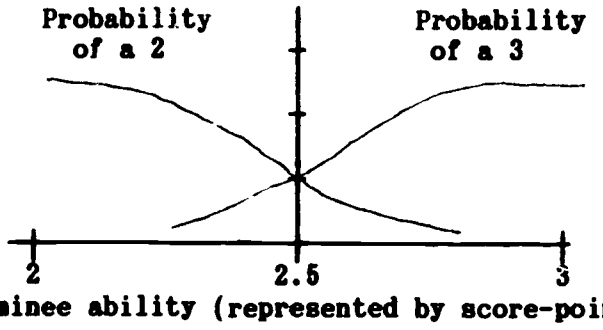

Figure 7. Probability of rating expected to be given to an examinee of given ability.

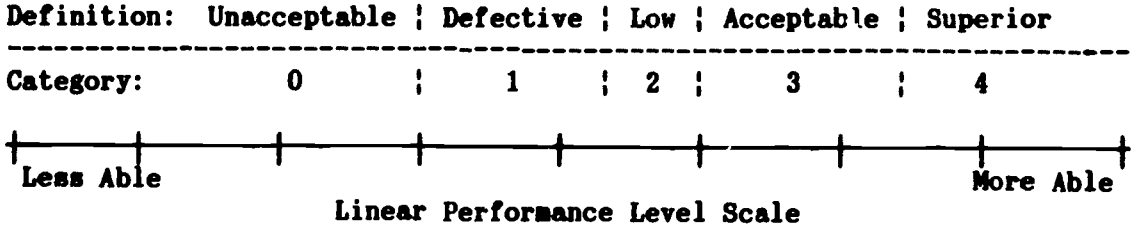

Figure 8. Relationship between category numbers and performance level.

III. Moving from reliability to objectivity: the nature of the rating scale.

The fact that the scale we have been discussing has categories numbered 0,1, 2,3,4 does not force the categories to represent equal increments in performance. In the next judging session, the examining board might decide to introduce a new category between previous categories 2 and 3, and then renumber the scale as 1,2,3,4,5,6. If the old scale were linear, then the new scale isn't, and vice-versa. The numbers assigned to categories are merely a convenience of labelling which enables an ordering of performance levels, but they are not a direct expression of the amount of performance each category represents. Figure 8 illustrates this in the context of a realistic rating scale. Category 0, "Unacceptable", represents all levels of performance below category 1, an infinite range, and similarly category 4,

"Superior" represents all levels of performance above category 3. Consequently, no matter how the categories are defined, it is impossible for all of them to represent equal ranges of performance, and thus it is impossible for a numerical scale made up of category numbers to be linear.

We have seen illustrated in Figure 7 that there is a stochastic element to the awarding of categories. If we extend this finding to Figure 8, it can be seen that category 2 represents such a narrow range of real performance, that a "true" or latent performance level on the threshold between a 1 and a 2 could well be rated not only as a 1 or a 2, but even as a 3. In fact, the stochastic nature of judge rating implies that whatever the examinee's performance level, there is some probability that the judge may award a rating in any of the categories, though, for a well designed scale, the category nearest the examinee's performance level has the highest probability. Figure 9 depicts what occurs. This stochastic behavior is what has caused the quest for reliability to fail, but it is this very behavior which provides the key to objectivity and so fairness.

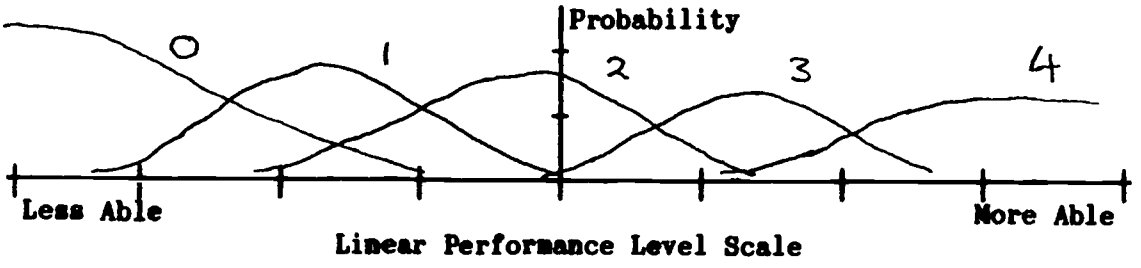

Figure 9. The probabilistic nature of the awarding of ratings by judges.

IV. The aim of the judging process.

For the examination described here, the ultimate goal of the judging process, from the viewpoint of an examining board, is not to determine some "true" rating on which ideal judges would agree, but rather to estimate the examinee's latent ability level, of which each judge's rating is a manifestation. This is the very essence of objectivity. In order to supersede the local particularities of the judging situation, each judge must be treated as though he has a unique severity, each examinee as though he has a unique ability, and each rating scale as though it has a unique formulation. This means that many interesting aspects of behavior must be regarded as incidental, and each rating considered to be the probabilistic result of only three interacting components: the severity of a judge, the ability of an examinee and the structure of the rating scale. With these assumptions, it is possible to obtain the outcome that the examining board desires, which is an estimate of the ability of each examinee, freed from the level of severity of the particular judges who happened to rate his performance and also freed from the arbitrary manner in which the categories of the rating scale have been defined. The more that incidental aspects of behavior are in evidence in the ratings, the more uncertainty there is in the estimates of the examinees' abilities, and the less confidence there is that the aim of the judging process has been realized in the judges' ratings.

Accurate measurement thus depends not on finding the one "ideal" judge but in discerning the intentions of the actual judges through the way in which they have replicated their behavior in all the ratings each has made. Consequently, judges cannot be assumed to be replications of one another, and to assert that examinees are sampled from a normal population is to predetermine the results of the examination. This indicates that judge training would be more productive if it were aimed towards fostering consistent behavior representative of the judge's intentions, rather than towards making the judge replicate a notional "ideal" judge.

		Judge F	
Categories:		2	3
Judge G	2	N22	N32
	3	N23	N33

Figure 10. Hypothetical counts of the awarding of categories by replications of two judges and the same examinee. Only instances when both judges have used categories 2 and/or 3 have been recorded.

		Judge F	
Categories:		2	3
Judge G	2	PG2*PF2	PG2*PF3
	3	PG3*PF2	PG3*PF3

Figure 11. Probabilities of the awarding of categories by two judges. Only instances when both judges have used categories 2 and/or 3 have been recorded. PF2 is the probability of Judge F awarding a 2, and PG2 is the probability of Judge G awarding a 2, and similarly for PF3 and PG3.

V. The probabilistic nature of ratings as a means to objective measurement.

It is the overlapping of category probabilities in Figure 9 that provide the means for objective measurement. For a judge of given severity, each performance level is determined uniquely by the probabilities of awarding the different categories. These probabilities are realized empirically in the data by the frequencies with which each judge awards each category. In principle, if two judges were to rate innumerable replications of the same examinee, the relative frequency with which they award the different categories would reveal their relative severity precisely and uniquely. Figure 10 summarizes an example of the counts of the ratings after a number of such replications.

However, the examining board's intention is that the judging situation be stable, so that, as the number of replications increase, the frequencies in Figure 10, when divided by the number of replications, approach the

underlying probabilities, PG2 that Judge G awards a 2 and PG3 that he awards a 3, and similarly PF2 and PF3 for Judge F. Since the two judges rate independently, the probabilities of their pairs of ratings are as shown in Figure 11.

The comparison of the severities of Judges F and G is based on the information in Figure 11 which contrasts their behavior. This is expressed by the probabilities in the cells which reflect their disagreement, the upper right and bottom left cells. The quantitative comparison itself is formed by the ratio (PG2*PF3)/(PG3*PF2). The top left and bottom right cells of Figure 11 give the probabilities of judge agreement, but they do not directly allow us to contrast the judges' behavior, and so do not reveal their relative severity. The probabilities themselves cannot be observed directly, even in theory, but they underlie the frequencies observed in the corresponding cells of Figure 10. Thus the quantitative comparison of Judges F and G is estimated by the ratio (N32/N23) of Figure 10.

This theoretical pair-wise comparison of judges could be continued across all pairs of judges, all pairs of categories, and all examinees. Further, a comparison of the abilities of each pair examinees could be made in the same way by considering the ratings awarded by numerous replications of judges of the same severity. In practice, however, the nature of the replications contained in the judges' ratings is not of numerous repetitions of the identical situation, but rather of numerous repetitions of the identical process. Each repetition contains some different subset of the parameters of the examination: judges' calibrations, examinee measures and rating scale calibrations. For objectivity, the values of these parameters must be regarded as fixed throughout the examination.

The nature of the rating scale, that is how far apart the categories boundaries are along the performance scale, is revealed by how the ratings given by judges of particular severity to examinees of particular ability are distributed among the categories. The severity of each judge is determined by his perception of overall examinee performance as revealed in the distribution of the ratings he awarded. The performance level of each examinee is determined by means of the performance levels implied in the ratings given by judges, each with a particular level of severity. Thus each examinee is represented by one measure of ability, as the examining board intended, and each judge by one measure of severity. Parameters are thus estimated jointly, but, once this process has been successfully accomplished, it no longer matters which judges rated which examinees.

VI. The objective measurement model.

The comparison of judges, presented in Figure 11, leads directly to the model necessary and sufficient for objective measurement of the examinees, which must obtain if the empirical set of ratings is to be statistically coherent, with all the parameters expressed on a linear scale. This measurement model, which is necessary for objectivity, was first proposed by Georg Rasch (1960/1980), and its extension to rating scales is given in Wright and Masters (1982). It has been further generalized to many-faceted judging situations (Linacre 1987).

The measurement model applicable to the examination described here is:

$$\log(P_{njk}/P_{njk-1}) = B_n - C_j - F_k$$

where

P_{njk} is the probability of examinee n being awarded by judge j a rating in category k

P_{njk-1} is the probability of examinee n being awarded by judge j a rating in category k-1

B_n is the ability of examinee n (the performance level on a linear scale which the examining board really wants)

C_j is the severity of judge j

F_k is the difficulty of the step from category k-1 to category k of the rating scale.

This model takes advantage of the fact that there is some probability of any judge rating any examinee in any category, as presented in Figure 9. This model, however, dictates the precise form of the probability curves necessary for objective measurement. The modelled forms of the curve must, and do, occur in practice when the judging process is truly in accord with the examining board's intentions that each examinee can be characterized by one ability parameter.

In the model equation, the logarithm of the ratios of probabilities, the "log-odds", is used to determine the ability of each examinee, the severity of each judge and the structure of the rating scale. The estimated measures and calibrations are obtained by fitting the actual ratings given by the judges into the framework determined by the model. This can be done by means of the technique of maximum likelihood which yields the estimated measures and calibrations which are most likely to produce the ratings that were awarded. These estimates are in "log-odds units" (logits) which form an interval scale and are equivalent to the inches or meters of physical science. The estimation procedure also provides standard errors, which show how accurately the measures have been determined. A further vital outcome are fit statistics which indicate whether the process of measuring the examinee's ability has been successful.

When the entire set of ratings is statistically coherent, then each rating cooperates in the simultaneous estimation of the parameters of the three facets within one overall framework, provided that sufficient overlap has been built into the judging plan to allow all judges and examinees to be integrated into a single global frame of reference. Neither complete data (every judge rating every examinee) nor complex judging plans (e.g. partial incomplete block designs) are required.

VII. Objective measurement for more complex judging situations.

The discussion in this paper has presented a simplified problem, as one example of many-faceted Rasch measurement. This theory can be applied to measurement in more complex situations. In principle, each new facet of a judging situation introduces into the general model equation its own set of

parameters. There is no requirement that every rating be formed out of the same combination of facets, only that the ratings form part of one overall design.

An example would be a certification examination in which each examinee is rated on one skill item by two judges from a panel of judges. This skill item is rated on a 5 point rating scale, and could be the performance of some laboratory procedure. Each candidate is also rated by another judge as to his success or failure on a second skill item, which could be the accuracy of his report of the outcome of the laboratory procedure. The judges are rotated so that each examinee is rated by three different judges, and each judge rates both skill items and is also paired with every other judge over the course of the judging session.

The measurement model for the first skill item, with 5 categories, thus becomes

$$\log(Pn1jk/Pn1jk-1) = Bn - D1 - Cj - F1k$$

where
- Pn1jk is the probability of examinee n's performance on the first skill item being awarded by judge j a rating of k
- Pnjk-1 is the probability of examinee n being awarded k-1
- Bn is the ability of examinee n
- D1 is the overall difficulty of skill item 1
- Cj is the severity of judge j
- F1k is the difficulty of the step from category k-1 to category k for skill item 1.

This first model is invoked simultaneously with the measurement model for the second dichotomous skill item:

$$\log(Pn2j1/Pn2j0) = Bn - D2 - Cj$$

where
- Pn2j1 is the probability of examinee n's performance on the second skill item being rated by judge j as successful
- Pn2j0 is the probability of examinee n being rated as failed, so that Pn2j1 + Pn2j0 = 1
- Bn is the ability of examinee n (same as for item 1)
- D2 is the difficulty of skill item 2
- Cj is the severity of judge j (same as for item 1)

The simultaneous application of these models would produce objective, linear, measures for each examinee, with their associated standard errors. The self-consistent behavior of the judges as well as the overall success of the measurement process could be verified by reference to well-defined fit statistics, in spite of the fact that this design includes very little duplicate judging and would not be amenable to most forms of analysis.

In situations in which examinees are rated on a number of items, duplicate ratings can be avoid entirely. This can be done by arranging for each

component part of an examinee's performance to be rated by a different judge of a judging team, with a judging plan which rotates judges across skill items and into different judging teams during the judging session.

VIII. Conclusion.

Attempting to determine an examinee's performance in terms of judge agreement on a "true" rating is seen to be a hopeless task. Even with judges of equal severity, their agreement, and hence reliability, will be affected by how close an examinee's performance is to a category boundary. Further, the arbitrary definition of the categories makes any attempt to use their numbers as the basis of arithmetical operations an exercise in dubious approximations.

Objectivity in examinations is obtained through a consideration of intention. Are examinee measures to be based on serendipitous numerical agreement in the ratings given by the judges, or are the examinee measures to be determined from the intentions of the judges as revealed through a consideration of the information contained in all their ratings ? If the intention of the examination board is to determine a measure for examinee on an interval scale amenable to arithmetical manipulation and generalizable beyond the particular details of the judging situation, then the many-faceted Rasch measurement model is the model required for such objectivity.

Bibliography:

Andrich, D. & Constable, E. Inter-judge reliability: is complete agreement among judges the ideal ? Paper presented at NCME, New Orleans. 1984.

Braun, H.I. Understanding Scoring Reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 1988, 13:1 pp. 1-18.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20:1 p.37-46.

Guilford, J.P. Psychometric Methods. (2nd Edn.) New York: Mc-Graw Hill. 1954.

Harper, A. E. Jr & Misra, V. S. Research on Examinations in India. National Council of Educational Research and Training, New Delhi, India. 1976

Linacre, J.M. An Extension of the Rasch Model to multi-faceted situations. Chicago: University of Chicago, Department of Education. 1987.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, 1960, and Chicago: University of Chicago Press. 1980.

Wright, B.D. & Masters, G.N. Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press. 1982.